J Autism Dev Disord
DOI 10.1007/s10803-007-0521-y

[illegible]

Audiovisual Processing in Children with and without Autism Spectrum Disorders

Elizabeth A. Mongillo · Julia R. Irwin · D. H. Whalen · Cheryl Klaiman · Alice S. Carter · Robert T. Schultz

© Springer Science+Business Media, LLC 2008

Abstract Fifteen children with autism spectrum disorders (ASD) and twenty-one children without ASD completed six perceptual tasks designed to characterize the nature of the audiovisual processing difficulties experienced by children with ASD. Children with ASD scored significantly lower than children without ASD on audiovisual tasks involving human faces and voices, but scored similarly to children without ASD on audiovisual tasks involving nonhuman stimuli (bouncing balls). Results suggest that children with ASD may use visual information for speech differently from children without ASD. Exploratory results support an inverse association between audiovisual speech processing capacities and social impairment in children with ASD.

Keywords Audiovisual processing · Visual influence · McGurk effect

Autism and related pervasive developmental disorders that fall into the overarching category of Autism Spectrum Disorders (ASD) (e.g. Asperger's Syndrome, PDD-NOS), typically have their onset within the first few years of life (Osterling and Dawson 1994; Volkmar and Lord 1998). Children with ASD typically experience difficulties with social reciprocity, language, and communication (Tager-Flusberg 1981, 1982; Volkmar and Lord 1998). There are numerous reports that children with ASD evidence impaired performance on a variety of speech perception tasks relative to mental age matched children without ASD. Reported impairments range from difficulties using visual information in a meaningful way in the presence of simultaneous auditory information to difficulties detecting correspondences between affect suggested by facial expressions versus affect suggested by tone of voice (de Gelder et al. 1991; Kikuchi and Koga 2001; Loveland et al. 1995; Massaro 1998; Massaro and Bosseler 2003; Williams et al. 2004).

There are two major classes of theories that describe how speech information is processed, those that suggest that speech perception is multimodal and those that suggest it is "amodal." According to multimodal accounts, incoming information about the speech signal is both auditory and visual in nature. Amodal accounts suggest the common underlying information about the speech signal is amodal and gestural in nature and that what the perceiver detects are the articulatory gestures, which are available from both the speaker's face and voice. Massaro's Fuzzy-Logical Model of Perception (Massaro and Bosseler 2003), an example of the former class of theories, suggests that incoming auditory and visual information from speech are evaluated in parallel and combined to create a meaningful (logical) audiovisual percept. Information from one modality may be weighted more heavily than information from another modality,

Electronic supplementary material The online version of this article (doi:10.1007/s10803-007-0521-y) contains supplementary material, which is available to authorized users.

E. A. Mongillo (✉) · A. S. Carter
Department of Psychology, University of Massachusetts Boston,
100 Morrissey Blvd., Boston, MA 02125-3393, USA
e-mail: elizabeth.mongillo@aya.yale.edu;
lizmongillo@gmail.com

J. R. Irwin · D. H. Whalen
Haskins Laboratories, New Haven, CT, USA

C. Klaiman · R. T. Schultz
Yale School of Medicine Child Study Center, Yale University,
New Haven, CT, USA

depending on the perceived value of each channel (auditory and visual) in reducing the ambiguity of the audiovisual stimulus presented (Massaro and Bosseler 2003). The latter class of theories including The Motor Theory of speech perception (Liberman et al. 1967; Liberman and Whalen 2000) and Direct Realist theories (Best 1995; Fowler and Smith 1986) argue that the speech signal is analyzed not in terms of separate auditory and visual components but in terms of shared underlying motor movements/gestures that specify the speech signal.

Typically developing individuals seem to be influenced perceptually by both the auditory and visual components of speech (Massaro 1984; Sumby and Pollack 1954). It is surprising that these effects occur in individuals with normal hearing who do not usually consider themselves to be lip-reading. The visual influence is both automatic and strong. On the FLMP account, this is because of life-long associations between the auditory and visual components of speech. On gestural accounts, it is due to the common event that is specified by each modality. Predicting the weight of each modality is a challenge for either account.

In our day-to-day life, auditory and visual components of speech are congruous (i.e. the lip-movements of the speaker and the sounds generated by his or her voice reflect the same event). Current technology allows us to present artificially manipulated stimuli in which the audio and visual stimulus components specify different syllables. Put another way, it is possible to present visual footage of a speaker mouthing one word or sound while simultaneously presenting audio footage of the same speaker uttering a different word or sound (e.g. showing lip movements that suggest the speaker is making the sound /a/ while simultaneously playing audio of the speaker making the sound /i/).

Mismatches between audio and visual tokens of the spoken vowels /a/, /i/ and /u/ are fairly easy to detect and our ability to do so develops as early as 4-months of age (Kuhl and Meltzoff 1984a, b). However, there are some instances of audiovisual mismatch in which the syllable that is "heard" by the listener is influenced by the syllable that has been simultaneously presented visually. This latter phenomenon, known as the McGurk Effect, occurs when the visual signal specifies the place of articulation of a particular stop, even though the acoustic signal is unambiguously a different stop (McGurk and MacDonald 1976). For instance, when the speaker's lip-movements for /gaga/ are paired with an auditory /baba/, some individuals will "hear" a /dada/ (McGurk and MacDonald 1976). This visual influence on heard speech has been shown in infants as young as 4-months of age and more consistently in infants 6-months of age (Desjardins and Werker 2004; Rosenblum et al. 1997).

For those susceptible to it, the McGurk Effect is extremely robust (Rosenblum et al. 1997), replicated under numerous conditions, and known to persist even if the visual information is degraded (e.g. the face of the speaker is asymmetrically scrambled, the chosen word is embedded in a sentence, or very brief visual stimuli are employed) (Green and Kuhl 1986; Heitanen et al. 2001; Irwin et al. 2006; MacDonald et al. 2000; Massaro and Cohen 1983; Sams et al. 1998; Summerfield 1979). There is preliminary evidence that it may not be limited to speech stimuli. When presented with discrepant audiovisual tokens of plucks and bows played on a cello, participants' perceptions of what they are hearing (e.g. a bow vs. a pluck) are significantly influenced by the visual information being presented (Fowler and Rosenblum 1990; Saldaña and Rosenblum 1993).

Some have suggested that children with ASD do poorly on McGurk tasks because of deficits in attending to multimodal information (de Gelder et al. 1991); others implicate impaired lip-reading skills (Williams et al. 2004). Consistent with lip-reading hypotheses, Massaro and Bosseler (2003) point out that children with ASD show more of a visual influence in their perceptions of mismatched audiovisual stimuli after intensive training in lip-reading skills (Massaro and Bosseler 2003; Williams et al. 2004). de Gelder et al. (1991), however, have found significant differences in audiovisual processing between mental-age-matched individuals with and without ASD, but no differences in lip-reading ability. Though children with ASD were able to use visual information to correctly determine what the speaker had said in visual-only conditions, they failed to use this information if auditory information was also present. Additionally, Irwin et al. (2006) have reported the ability to lip-read does not automatically result in an ability to integrate and that lip-reading skills do not correlate with use of visual information in typical perceivers.

The difficulty that children with ASD demonstrate on McGurk tasks may be related to an underlying difficulty in attending to and using information from the face. Children with ASD often have trouble recognizing emotions and detecting inter-modal correspondence of facial and vocal affect (Kikuchi and Koga 2001; Loveland et al. 1995). Furthermore, they show abnormal brain activation in response to faces. While children without ASD show strong activation of the lateral fusiform gyrus during facial processing tasks, children with ASD show reduced levels of activation in their fusiform gyri (Schultz et al. 2000), and level of activity predicts accuracy on face perceptual tasks (Schultz et al. 2005).

One possibility that may account for these seemingly contradictory findings from the speech perception and face perception literature is that children with ASD may be using a parts vs. whole type of processing. Proponents of the Central Coherence Theory (Frith and Happe 1994) argue that autism may involve an impaired ability to perceive central coherence from individual features of a stimulus, such as a face. This theory is supported by recent evidence that individuals with ASD have difficulties both visually integrating objects to make a coherent scene and spontaneously detecting context-inappropriate objects (Jolliffe and Baron-Cohen 2001).

Difficulties on speech tasks like the McGurk may also be related to the “social” (human) aspects of the task. Dawson et al. (1998) have reported that children with autism have significantly more difficulty visually orienting to “social” stimuli (e.g. their name being called, sound of hands clapping) than children without ASD and children with mental retardation, but only slightly more difficulty visually orienting to “nonsocial” (nonhuman) stimuli (e.g. musical toy, sound of a rattle). Oftentimes, children with autism fail to orient to social stimuli at all, and when they do orient to social stimuli, their response tends to be delayed (Dawson et al. 1998).

There is significant disagreement over the nature and source of speech perception difficulties in children with ASD. Dawson et al.’s findings led us to suspect that children with ASD would evidence more pronounced difficulties relative to control sample (CS) children on audiovisual processing tasks involving “human” audio and visual stimuli (e.g. faces and voices) than tasks involving “nonhuman” stimuli (e.g. bouncing balls). The present study further explores differences in the use of visual information between children with and without ASD, and it begins to address whether children with ASD perform typically on similar kinds of tasks with nonhuman stimuli. A McGurk task (McGurk and MacDonald 1976) was used to assess audiovisual processing along with several control conditions that were designed to determine whether children with ASD could process audio and visual information if the task involved an explicit comparison of the two modalities. The study also included several different types of audiovisual mismatches to determine whether the children with ASD possessed a general processing difficulty, or whether processing difficulties were more pronounced with stimuli involving human faces and voices than with stimuli involving objects (e.g. bouncing balls).

Methods

Participants

This study included 15 children with autism spectrum disorders, 2 females and 13 males (mean age = 13.73) age range 8–19 years, and 21 controls without autism spectrum disorders, 10 females and 11 males (mean age = 13.44 years), age range 11–19 years.

All participants with an ASD diagnosis were recruited through projects in a Collaborative Programs of Excellence in Autism (CPEA) grant based at the Yale Child Study Center. Participants in the control sample (CS) were recruited from the North Haven Middle School, through parents employed as staff at the Yale Child Study Center, and from area group homes. Children from group homes were included to allow matching of the mean IQ and IQ range to the ASD sample.

All participants with an ASD diagnosis were assessed with state of the art diagnostic procedures, including the Autism Diagnostic Observation Schedule (ADOS; Lord et al. 1999), the Autism Diagnostic Interview-Revised (ADI-R; Lord et al. 1994), and met criteria for ASD based on expert clinician consensus. Nine of the children in the ASD group met ADOS and ADI-R criteria for autism, while six showed more typical language by age 3 and met criteria for Asperger’s syndrome. Social Reciprocity Scale (SRS) (Constantino et al. 2000; Constantino and Todd 2000) scores were also obtained for 11 of the 15 participants with ASD and were used to assess level of social impairment.

IQ information was collected on all 15 of the children with autism spectrum disorders (mean verbal IQ = 99, range: 54–134; mean performance IQ = 93.2, range: 71–129; mean full scale IQ = 96.13, range: 67–134). IQ information was also collected on 19 of the children in the control group (mean verbal IQ = 103.61, range: 60–137; mean performance IQ = 100, range: 53–126; mean full scale IQ = 102.11, range: 53–136). Some participants were given the full WISC-III and others were given the briefer WASI, which includes 4 subscales of the WISC-III.

All parents of children in the control group were given a short questionnaire asking them to report whether their child had any known learning disabilities, Attention-Deficit Hyperactivity Disorder (ADHD), psychiatric conditions, social disabilities, or first or second degree relatives with an autism spectrum diagnosis. Children who met criteria for any of these disorders and those whose parents reported any of these problems were excluded from participating in the study. Three of the controls were unable to return for the follow-up IQ assessments, but parental reports gave us no reason to believe that their IQs would not have been within the normal range. An additional three of the controls had FSIQ scores <70. These individuals had a nonspecific or otherwise unidentified form of mental retardation, and were recruited from area group homes so as to allow matching of the mean IQ and IQ range to the ASD sample. Low IQ participants had no known conditions other than MR and had not been diagnosed with autism or any of the exclusionary conditions listed on the questionnaire.

Participants were matched at the group rather than the individual level because ultimately, the aim of the study was to detect group rather than individual differences. Participants were matched in mean age and IQ but were not matched for in terms of gender. Though there is some evidence of gender differences in ability to lip-read words, research to date suggests that no such gender differences have been found in ability to lip-read shorter speech stimuli (Irwin et al. 2006).

Consent was obtained and the study was approved by the Yale IRB. Participants were given a $15 gift certificate to a bookstore or music store as compensation for their time participating in this study.

Procedures

Participants were asked to complete 6 perceptual tasks, a purely visual Male/Female Face Classification Task, and 5 audiovisual tasks: the McGurk Task, and the Gender, Vowel, Ball Size and Ball Composition Match/Mismatch Tasks. Descriptions of the tasks can be found in the measures section. Each of the 6 tasks were presented on a Macintosh laptop computer using PsyScope (Cohen et al. 1993). Testing was conducted in a small quiet room and took approximately 40–45 min to complete. Every child received the Male/Female Face Classification Task first, as a warm-up exercise. The five audiovisual tasks were then presented in a random order. Each consisted of a familiarization segment containing non-scored items used to get participants familiar with the format of the test, followed by a test segment, in which the computer kept track of correct vs. incorrect responses. After practice, no feedback was given as to whether test responses were correct or incorrect on any of the 6 tasks.

To investigate our main hypothesis, we compared the performance of the ASD group with that of the CS group on each of the perceptual tasks. On the Male/Female Face Classification Task, the performance score was calculated with the following algorithm: (# of correctly identified male faces + # of correctly identified female faces)/total number of face stimuli presented. On the McGurk Task, the performance score was calculated with the algorithm: (# of correct responses to catch trials + # responses indicating visual influence)/total number of trials. On the 4 remaining AV match–mismatch tasks the performance score was calculated using the algorithm: (# of correctly identified AV matches + # of correctly identified AV mismatches)/total number of AV pairs presented.

Measures

Autism Diagnostic Observation Schedule (ADOS) (Lord et al. 1999). The ADOS is a semi-structured, interactive observation schedule designed to assess social and communicative functioning in individuals who may have an autism spectrum disorder. A standardized diagnostic algorithm can be computed, composed of a subset of rated social and communicative behaviors, and consistent with autism criteria in DSM-IV/ICD-10.

Autism Diagnostic Interview-Revised (ADI-R) (Lord et al. 1994). The ADI-R is a structured interview with a parent on early development pertinent to autism symptomatology, including the presence of communication skills prior to age 3 years.

Social Reciprocity Scale (SRS) (Constantino et al. 2000; Constantino and Todd 2000). The SRS is a 65-item questionnaire completed by the child's parents regarding their child's social interactions. It provides a very sensitive measure of social reciprocal behavior, which is uncorrelated with IQ, but strongly correlated with scores on the social deficit scale of the ADI-R (r = 0.80). The SRS has high sensitivity and specificity in distinguishing children with pervasive developmental disorders such as autism from children with other psychiatric disorders.

Male/Female Face Classification Task. Thirty black and white photographs of either a male or a female face with hair and other extraneous identifying features removed were presented to participants. The participants were presented with faces one by one, in a random order, on a computer screen. They were given as much time as they needed to look at each face and decide via a 2-choice button press whether it was male or female.

The McGurk Task. Matched and mismatched audiovisual renditions of /ba/, /da/, /va/, and[1] /ða/ were created using Adobe Premiere 6.0. The auditory and visual dubbed signals were created by aligning the onset of the new acoustic signal with the original acoustic signal at point of release. The timing of auditory and visual signals was matched down to the accuracy of a single frame (33 ms), ensuring that there were no obvious temporal lags and all tokens were carefully examined for auditory and visual clarity. All speech sounds in this task were produced by a single female speaker, whose utterances have been shown to effectively elicit the McGurk Effect in typically developing adults (Irwin et al. 2006). The audiovisual stimuli were either congruent or incongruent. The congruent stimuli, simultaneous presentation of audio of the speaker saying /ba/ and video of the speaker saying /ba/ (audiovisual matches), were the original productions of the speaker. The incongruent stimuli were created by digitally recombining audio of the speaker saying /ba/ with video of the speaker saying either /va/, /da/, or /ða/ (audiovisual mismatches).

[1] This token is a voiced dental fricative; the consonant is the initial consonant of "the".

In the familiarization phase, participants were presented with matched, original footage, audiovisual renditions of /ba/, /da/, /va/, and /ða/ and were asked to indicate by pressing one of two buttons, whether or not the speaker said /ba/. In the test phase, participants were presented with 30 matching and mismatching audiovisual tokens in a random order and asked to press one of two keys on the computer to indicate whether they heard the speaker say /ba/ or something else. In nearly all test stimuli, the audio presented was /ba/. For half of the /ba/ audio stimuli, the simultaneous video was also /ba/. For the other half of the /ba/ audio stimuli, the simultaneous video was something other than /ba/—either /da/, /va/, or /ða/.

There were also a few catch trials in which the audio was not /ba/. The number of catch trials on the McGurk Task was determined by the number of visual stimuli to be paired with audio /ba/ in the test trials. The stimuli presented in the catch trials were the same as those presented in the familiarization trials and consisted of matched audiovisual tokens of audio /va/–visual /va/, audio /da/–visual /da/, and audio /ða/–visual /ða/. Each of the correctly matched combinations was presented once. Items on the McGurk Task were scored as correct when the response was consistent with the visual input.

AV Match–Mismatch Task. The stimuli for the Gender and Vowel AV Match–Mismatch Tasks were created by recording, digitizing and re-combining audiovisual tokens of a number of male and female speakers saying /a/, /i/, and /u/. The stimuli for the Ball Size and Ball Composition AV tasks were created in a similar manner by recording, digitizing and recombining audiovisual tokens of several types of balls bouncing on a metal surface.

All of the editing was accomplished with the program Adobe Premiere 6.0. As in the creation of the McGurk stimuli, the auditory and visual dubbed signals for the AV Match–Mismatch stimuli were created by aligning the onset of the new acoustic signal with the original acoustic signal at point of release. The timing of auditory and visual signals was matched down to the accuracy of a single frame (33 ms), ensuring that there were no obvious temporal lags. All tokens created were first piloted with a group of typically developing adults in order to eliminate any tokens that were lacking in visual or auditory clarity. For each task, an even number of congruent (matched) and incongruent (mismatched) trials were selected from among the remaining tokens. Given that different tasks required different types of combinations and that some combinations had to be eliminated due to problems with auditory or visual clarity, the number of test trials was not consistent across tasks.

The test trials on each of the AV Match–Mismatch Tasks were preceded by a familiarization phase in which participants were prompted to practice pressing the “match” key when presented with examples of congruent audiovisual tokens (e.g. female face speaking in a female voice) and to press the “mismatch” key when presented with examples of incongruent audiovisual tokens (e.g. female face speaking in a male voice). In the test phase, participants were presented with new congruent and incongruent audiovisual tokens and asked to classify them as matches or mismatches by pressing the “match” or “mismatch” key on the computer. Each test combination was presented only once. Given that different tasks required different types of combinations and given that not every possible combination made it past the pilot phase, the number of test trials was not consistent across tasks.

AV Gender Match–Mismatch Task. Participants were presented with 24 audiovisual tokens, half of which were congruent (designed to register as “matches”) and half of which were incongruent (designed to register as “mismatches”). In the congruent pairs, the speaker’s original voice was replaced by the voice of another speaker of the same sex, and in the incongruent pairs, the speaker’s original voice was replaced by the voice of another speaker of the opposite sex. A speaker was never paired with an instance of his or her own voice in the familiarization phase or the test trials, and the instructions given to participants were very clear that the task was to identify matches and mismatches based on gender only, independent of speaker. The face and voice of the speaker were always congruent in terms of vowel sound presented regardless of whether the stimulus was congruent or incongruent in terms of speaker gender.

AV Vowel Match–Mismatch Task. Participants were presented with 47 audiovisual tokens, approximately half of which were congruent and half of which were incongruent.[2] In the congruent pairs, the speaker’s original voice was replaced by a different instance of his or her voice uttering the same vowel sound, while in the incongruent pairs, the speaker’s original voice was replaced by a different instance of his or her voice uttering a different vowel sound. For example, some stimuli consisted of a simultaneous presentation of audio of the speaker saying /a/ and video of the speaker saying /a/. The remaining stimuli were made with /i/ as the audio paired with visual /i/, /a/ or /u/, and with /u/ as the audio paired with visual /u/, /a/ or /i/.

AV Ball Size Match–Mismatch Task. Participants were presented with 24 audiovisual tokens, half of which were congruent and half of which were incongruent. In the congruent pairs, the ball’s original bouncing sound was replaced by the sound from another instance of that same

[2] The total number of trials presented on the Vowel Task should have been 48, but a computer glitch resulted in the presentation of only 47 trials. The missing trial was one of a number of a–v stimuli featuring a combination of audio /a/ and visual /a/. All other combinations of audio /a/ and visual /a/ were presented as intended.

ball bouncing, while in the incongruent pairs, it was replaced by the sound from a different sized ball bouncing.

AV Ball Composition Match–Mismatch Task. Participants were presented with 36 audiovisual tokens, half of which were congruent and half of which were incongruent. In the congruent pairs, the ball's original bouncing sound was replaced by the sound from another instance of that same ball bouncing, while in the incongruent pairs it was replaced by the sound from a different type of ball bouncing (e.g. the audio of a ping-pong ball bouncing with the visual image of a rubber ball bouncing).

Results

Multivariate Analyses of Variance (MANOVAs) revealed no significant gender differences in task performance among the children in the control sample on the six perceptual tasks (all p-values > 0.4), and no gender differences on age or IQ scores (all p-values > 0.2). Thus, the 10 female and 11 male controls were combined ($n = 21$) for all further analyses. Mean age and IQ for the control sample (CS) and the ASD group are reported in Table 1. A MANOVA, with group (ASD vs. CS) as a fixed factor and age, full scale IQ, verbal IQ and performance IQ as dependent variables revealed no significant group differences in age or IQ (all p-values > 0.3).

To test our primary hypothesis that there would be a task by group interaction, a Univariate Analysis of Variance (ANOVA) was conducted with group (ASD vs. CS) and task type[3] (Male/Female Face Classification vs. AV Gender Match–Mismatch vs. AV Vowel Match–Mismatch vs. McGurk vs. AV Ball Composition Match–Mismatch vs. AV Ball Size Match–Mismatch) entered as fixed factors and Task performance entered as the dependent variable. A more stringent criterion for significance ($p < 0.001$) was used in order to control for Type 1 error. There were significant main effects for both group F[1,203] = 22.98, $p < 0.001$, and task F[5,203] = 24.22, $p < 0.001$, and a significant task × group interaction F[11,203] = 16.61, $p < 0.001$.

To further explore the task × group interaction and identify which tasks were driving it, 6 separate univariate ANOVAs were conducted, each of which included group (ASD vs. CS) as the fixed factor and performance on one of the six perceptual tasks as the dependent variable. There was a main effect of group on three of the six tasks: the Male/Female Face Classification Task (F[1,34] = 13.41, $\eta^2 = 0.28$, $p < 0.001$), the McGurk Task (F[1,34] = 19.01, $\eta^2 = 0.36$, $p < 0.0001$), and the AV Vowel Match/Mismatch Task (F[1,34] = 15.89, $\eta^2 = 0.32$, $p < 0.0001$), with the ASD group having performed significantly less accurately on each task. For the McGurk task, this meant that the participants with an ASD diagnosis experienced less visual influence. Group means on the six perceptual tasks[4] are listed in Table 2.

Table 1 Sample characterization data

Measure	CS	ASD
Full scale IQ	102.1 [19.0]	96.1 [21.8]
Verbal IQ	103.6 [18.2]	99 [24.8]
Performance IQ	100 [20.2]	93.2 [18.0]
Age	13.4 [2.8]	13.7 [3.9]
ADOS social int. sum	NA	16.7 [6.4]
ADOS social algorithm	NA	9.9 [3.6]
SRS social reciprocity	NA	101.7 [13.0]

CS = Control Sample Group ($n = 19$), 2 children excluded because did have IQ scores; ASD = Autism Spectrum Disorder Group ($n = 15$); Standard deviations are listed in [ ] beside group means

The group results may have been mediated by specific stimulus types, since each experimental task was composed of more than one type of stimulus. To further explore this, ANOVAs were conducted with two independent variables—group (ASD vs. CS) and stimulus type (e.g. male or female face in the face task, audiovisual matches vs. mismatches in the other 2 tasks). There was a main effect of stimulus type on the Male/Female Face Task (F[1,68] = 5.91, $p < 0.02$). Both groups had more difficulty correctly labeling female faces than male faces. On the McGurk Task there was both a main effect of stimulus type (F[1,176] = 46.28, $p < 0.001$), and a significant interaction between stimulus type and group (F[1,176] = 22.82, $p < 0.001$). While both groups performed similarly on trials involving matched stimuli (visual + auditory /ba/, /da/, /ða/, and /va/), children in the ASD group performed significantly differently from children in the CS group on trials involving the mismatched stimuli (visual /da/, /va/, and /ða/ + auditory /ba/). As shown in Table 3, children with ASD demonstrated less visual influence on the following visual-auditory matching conditions: ba–da (F[1,34] = 9.80, $\eta^2 = 0.22$, $p < 0.005$); ba–ða (F[1,34] = 26.6, $\eta^2 = 0.44$, $p < 0.0001$); and ba–va

[3] The Male/Female Face Classification Task was treated as separate from the AV Gender Match–Mismatch Task. Performance on the former was not co-varied out when analyzing performance on the latter, because the two tasks employed entirely different stimuli. The gender of the faces used in the AV Gender Match–Mismatch task was much more easily identifiable than the gender of the faces in the Male/Female Face Classification Task, because the former included color images in which the entire head, hair, etc. were clearly visible and the latter included black & white images in which the hair had been cropped out.

[4] The tasks are referred to collectively as "perceptual" rather than "cross-modal" because only five of the six were cross-modal. The Male Female Face Classification Task was unimodal, as it included only a visual component and no auditory component.

Table 2 Mean task performance by group

Task	CS	ASD	F-values	η^2
Male/Female Face (visual only)	87.7 [6.9]	76 [12.1]	13.41*	0.28
AV Gender (audiovisual)	86 [12.1]	75.4 [21.4]	3.23	0.09
McGurk (audiovisual)	94.9 [8.1]	78.3 [14.5]	19.01*	0.36
AV Vowel (audiovisual)	94 [6.2]	78.5 [16.3]	15.89*	0.32
AV Ball Composition (audiovisual)	65.3 [13.9]	67.2 [13.0]	0.18	<0.01
AV Ball Size (audiovisual)	60.9 [10.3]	62.9 [12.6]	0.27	<0.01

CS = Control Sample Group (n = 21); ASD = Autism Spectrum Disorder Group (n = 15); Values reported in CS and ASD columns correspond to group means (% correct); Standard deviations are listed in [ ] beside group means; Significant differences are starred (*)

Table 3 McGurk performance by group and blend type

Task	CS	ASD	F-values	η^2
ba–da (a–v mismatch)	92.9 [14.0]	68.3 [32.0]	9.8*	0.22
ba–ða (a–v mismatch)	91.7 [16.5]	48.3 [33.4]	26.6*	0.44
ba–va (a–v mismatch)	86.9 [23.2]	45.0 [40.3]	15.58*	0.31
ba–ba (a–v match)	98.1 [5.8]	96.1 [10.9]	0.47	0.01
Catch trials (a–v match)	96.8 [8.5]	91.1 [18.8]	1.53	0.04

CS = Control Sample Group (n = 21); ASD = Autism Spectrum Disorder Group (n = 15); Values reported in CS and ASD columns correspond to group means (% correct); Standard deviations are listed in [ ] beside group means; Catch trials included the following matches: da–da, ða–ða va–va; Significant differences are starred (*)

(F[1,34] = 15.58, $\eta^2 = 0.31$, $p < 0.0001$). Within the ASD group, visual influence on /ba/–/da/ trials was significantly higher than on the /ba/–/ða/ or /ba/–/va/ trials (t = 8.27, $p < 0.001$). There were no main or interaction effects on the AV Vowel Match/Mismatch Tasks.

To examine whether performance on the McGurk task was correlated with performance on any of the other 5 tasks, Pearson correlations were computed for each group (CS and ASD). For each group, correlations between performance on the McGurk Task, the AV Gender Match–Mismatch Task, and the AV Vowel Match–Mismatch Task were significant at the 0.01 level (all r-values ≥ 0.6). Correlations between speech tasks were expected based on findings of other studies (Surprenant and Watson 2001). Exploratory analyses revealed that scores on the McGurk Task were also inversely correlated with SRS score (r = −0.489, $p < 0.01$), but were not significantly correlated with verbal, performance or full scale IQ.

Discussion

In this study, we attempted to determine whether and under what conditions children with ASD differ from mental age matched children without ASD in the use of visual information for speech. Children with ASD scored significantly lower than children without ASD on 3 of the 4 tasks involving human faces and voices (Male/Female Face Classification Task, McGurk Task and AV Vowel Match–Mismatch Task), yet scored similarly to children without ASD on both tasks involving non-human stimuli (AV Ball Size & Ball Composition Match–Mismatch Tasks).

Our initial hypotheses regarding the McGurk effect were largely supported by the data. Children in the ASD group exhibited a much weaker McGurk effect than children without ASD, supporting previous findings that children with autism are less influenced by visual speech information (de Gelder et al. 1991; Massaro 1998; Massaro and Bosseler 2003). Children in the ASD group performed at chance and showed very little visual influence on the McGurk audiovisual mismatches as a whole (54%), though they were slightly above chance with the ba–da mismatches (68%) and at chance with the ba–ða (48%) and ba–va (45%) mismatches. It is very interesting that children with autism seem to be more susceptible to the /ba/–/da/ combination than the other two mismatches. What is even more interesting is that they are less susceptible to /ba/–/va/, which is the strongest in typically developing controls controls (Irwin et al. 2006; Rosenblum and Saldaña 1996). Both of these differences are intriguing and will require further study.

Children with ASD may experience less of a McGurk effect because they pay less attention to the face in general (Klin et al. 2002). A replication of Saldaña and Rosenblum (1993) on a population with autism would better clarify what is happening. If children with ASD experienced a McGurk effect comparable to that of children without ASD with audiovisual mismatches of plucks and bows on a cello, it would indicate that their audiovisual processing disability is primarily social in nature. Given that children with and without ASD performed similarly on tasks involving nonhuman stimuli like bouncing balls, but differently on tasks involving human faces and voices, it is likely that children with and without ASD might perform similarly on a McGurk task comprised of images and sounds made by an inanimate object such as a cello.

Exploratory results support an inverse association between audiovisual speech processing capacities and social impairment in children with ASD. Poor social reciprocity skills and poor speech processing skills may be linked. From early childhood, children with ASD preferentially attend to nonsocial vs. social stimuli. For example, typically developing 5-year-olds prefer to listen to their mother's voice (speech), while 5-year-olds with ASD prefer to listen to non-speech noise (Klin 1991, 1992). Given their early preferences for the nonsocial over the social, it is not surprising that children with ASD would be poorer than children without ASD at detecting audiovisual matches and mismatches with face and voice stimuli but perform similarly to children without ASD with nonsocial stimuli (e.g. bouncing balls).

Studies of typically developing infants have shown that a listening preference for speech over non-speech sounds normally emerges in the first year of life. Auditory preferences for speech versus non-speech stimuli have been systematically studied in 4-month-olds, 5–8-month-olds, and 9-month-olds. Four-month-olds preferentially attend to a female voice over white noise (Colombo and Bundy 1981). Five to eight-month-olds smile and vocalize more when presented with human faces and sounds, e.g. other infants vocalizing, than with nonhuman faces and sounds, e.g. dolls producing non-speech noises (Legerstee et al. 1998). The preference for speech over non-speech continues to exist in 9-month-olds, and even extends into the musical domain. Infants presented with instrumental and vocal music matched on pitch, rhythm and amplitude, have a strong preference for the vocal music (Glenn et al. 1981).

The ability to integrate the visual and auditory components of speech appears to develop early as well, as many developing infants demonstrate the McGurk Effect (Desjardin and Werker 2004; Rosenblum et al. 1997). Though not all infants experience the McGurk by 4–5 months of age (Desjardin and Werker 2004), this ability may nonetheless be an important part of typical speech development, which may put children with autism at risk for language delays or may partially explain the language difficulties in this population.

Children without ASD also have an early appearing preference for the human face (Cassia et al. 2004; Johnson et al. 1991) in addition to speech sounds. Therefore, it is not surprising that they do very well on audiovisual match–mismatch tasks involving these types of stimuli. Infants with ASD do not show typical attentional preferences to faces and eyes (Baron-Cohen 1998). While children with and without ASD have daily exposure to human faces and speech sounds, children with ASD do not take the same interest in them as children without ASD.

This study has provided some preliminary data regarding the ways in which children with and without ASD perform on a variety of audiovisual tasks. However, it is important to point out that there are some limitations. Given that some of the stimuli were being piloted for the first time, there was a lack of balance in number of trials across tasks. There was also a lack of gender balance and number of participants across groups, though it is unclear whether these imbalances have practical import.

Given that the majority of participants with ASD were recruited as part of an ongoing study of high functioning individuals the results may not be generalizable to lower functioning individuals with ASD. The data are also limited in that the tasks were not designed in such a way so as to allow us to directly measure and distinguish between reduced overall attention to the face vs. impaired lip-reading abilities or to distinguish between the Fuzzy Logical Model, Motor Theory and Direct Realist accounts. These are all important future directions to pursue.

Given that many speech vs. non-speech preferences are present in early childhood and infancy, differential performance between children with and without ASD on audiovisual processing tasks may be present much earlier in development as well. We are currently looking for such differences in younger children in the hope of identifying further criteria by which to identify infants at risk for autism.

Acknowledgments Support for this research came from a grant (U19-HD35482) from the NICHD, as part of the Collaborative Programs of Excellence in Autism (CPEA). Drs. Irwin and Whalen were supported by NIH grant HD-01994. We also acknowledge our gratitude for the diagnostic assessments provide by Ami Klin, Fred Volkmar and Diane Goudreau, the ASD participant recruitment by Tammy Babitz, and the careful proofreading of our final draft provided by Roxanne Mongillo. Finally, we wish to thank the children and families who gave their time so generously to facilitate this research project.

References

Baron-Cohen, S. (1998). Does the study of autism justify minimalist innate modularity? *Learning & Individual Differences, 10*(3), 179–191.

Best, C. T. (1995). A direct realist perspective on cross-language speech perception. In W. Strange & J. J. Jenkins (Eds.), *Cross-langauge speech perception* (pp. 171–204). Timonium, MD: York Press.

Cassia, V. M., Turati, C., & Simion, F. (2004). Can a nonspecific bias toward top-heavy patterns explain newborns' face preference? *Psychological Science, 15*(6), 379–383.

Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments, and Computers, 25*(2), 257–271.

Colombo, J., & Bundy, R. S. (1981). A method for the measurement of infant auditory selectivity. *Infant Behavior & Development, 4*, 219–223.

Constantino, J. N., Przybeck, T., Friesen, D., & Todd, R. D. (2000). Reciprocal social behavior in children with and without pervasive developmental disorder. *Journal of Developmental & Behavioral Pediatrics, 21*(1), 2–11.

Constantino, J. N., & Todd, R. D. (2000). Genetic structure of reciprocal social behavior. *American Journal of Psychiatry, 157*(12), 2043–2044.

Dawson, G., Meltzoff, A. N., Osterling, J., Rinaldi, J., & Brown, E. (1998). Children with autism fail to orient to naturally occurring social stimuli. *Journal of Autism & Developmental Disorders, 28*(6), 479–485.

de Gelder, B., Vroomen, J., & Van der Heide, L. (1991). Face Recognition and Lip-reading in Autism. *European Journal of Experimental Psychology, 3*(1), 69–86.

Desjardins, R. N., & Werker, J. F. (2004). Is the integration of heard and seen speech mandatory for infants? *Developmental Psychobiology, 45*(4), 187–203.

Fowler, C. A., & Rosenblum, L. D. (1990). Duplex perception: A comparison of monosyllables and slamming doors. *Journal of Experimental Psychology: Human Perception and Performance, 16*, 742–754.

Fowler, C. A., & Smith, M. (1986). Speech perception as "vector analysis": An Approach to the problems of segmentation and invariance. In J. Perkell & D. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 123–136). Hillsdale, NJ: Lawrence Erlbaum Associates.

Frith, U., & Happe, F. (1994). Autism: Beyond "theory of mind'. *Cognition, 50*(1–3), 115–132.

Glenn, S. M., Cunningham, C. C., & Joyce, P. F. (1981). A study of auditory preferences in nonhandicapped infants and infants with Down's syndrome. *Child Development, 52*(4), 1303–1307.

Green, K. P., & Kuhl, P. K. (1986). The Role of Visual Information from a Talker's Face in Processing of Place and Manner Features in Speech. *Journal of the Acoustical Society of America, 80*(S–63).

Heitanen, J. K., Manninen, P., Sams, M., & Surakka, V. (2001). Does audiovisual speech perception use information about facial configuration? *European Journal of Cognitive Psychology, 13*(3), 395–407.

Irwin, J. R., Whalen, D. H., & Fowler, C. A. (2006). A sex difference in audiovisual integration of speech. *Perception and Psychophysics, 68*, 582–592.

Johnson, M. H., Dziurawiec, S., Ellis, H., & Morton, J. (1991). Newborns' preferential tracking of face-like stimuli and its subsequent decline. *Cognition, 40*(1–2), 1–19.

Jolliffe, T., & Baron-Cohen, S. (2001). A test of central coherence theory: Can adults with high-functioning autism or Asperger syndrome integrate objects in context? *Visual Cognition, 8*(1), 67–101.

Kikuchi, T., & Koga, S. (2001). Recognition of others' and own facial expressions and production of facial expression: Children and adults with autism. *Japanese Journal of Special Education, 39*(2), 21–29.

Klin, A. (1991). Young autistic children's listening preferences in regard to speech: A possible characteristic of the symptom of social withdrawal. *Journal of Autism and Developmental Disorders, 21*, 29–42.

Klin, A. (1992). Listening preferences in regard to speech in four children with developmental disabilities. *Journal of Child Psychology and Psychiatry, 33*, 763–769.

Klin, A., Jones, W., Schultz, R., Volkmar, F., & Cohen, D. (2002). Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Archives of General Psychiatry, 59*(9), 809–816.

Kuhl, P. K., & Meltzoff, A. N. (1984a). *Infants' representations of events: Studies in imitation, cross-modal perception, and categorization*. Paper presented at the fourth international conference on infant studies, New York.

Kuhl, P. K., & Meltzoff, A. N. (1984b). The intermodal representation of speech in infants. *Infant Behavior and Development, 7*, 361–381.

Legerstee, M., Anderson, D., & Schaffer, A. (1998). Five- and eight-month-old infants recognize their faces and voices as familiar social stimuli. *Child Development, 69*, 37–50.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74*, 431–461.

Liberman, A. M., & Whalen, D. H. (2000). On the relation of speech to language. *Trends in Cognitive Science, 4*, 187–196.

Lord, C., Rutter, M., DiLavore, P. C., & Risi, S. (1999). *Autism Diagnostic Observation Schedule-WPS (ADOS-WPS)*. Los Angeles, CA: Western Psychological Services.

Lord, C., Rutter, M., & Le Couteur, A. (1994). Autism Diagnostic Interview-Revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism & Developmental Disorders, 24*(5), 659–685.

Loveland, K. A., Tunali-Kotoski, B., Chen, R., & Brelsford, K. A. (1995). Intermodal perception of affect in persons with autism or down syndrome. *Development & Psychopathology, 7*(3), 409–418.

MacDonald, J., Andersen, S., & Bachman, T. (2000). Hearing by eye: How much spatial degradation can be tolerated? *Perception, 29*, 1155–1166.

Massaro, D. W. (1984). Children's perception of visual and auditory speech. *Child Development, 55*, 1777–1788.

Massaro, D. W. (1998). *Perceiving talking faces: From Speech perception to a behavioral principle*. Cambridge, MA: MIT Press.

Massaro, D. W., & Bosseler, A. (2003). Perceiving speech by ear and eye: Multimodal integration by children with autism. *Journal on Developmental and Learning Disorders, 7*, 111–144.

Massaro, D. W., & Cohen, M. M. (1983). Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 9*, 753–771.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746–748.

Osterling, J., & Dawson, G. (1994). Early recognition of children with autism: A study of first birthday home videotapes. *Journal of Autism & Developmental Disorders, 24*, 247–257.

Rosenblum, L. D., & Saldaña, H. M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 22*(2), 318–331.

Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Perception & Psychophysics, 59*(3), 347–357.

Saldaña, H. M., & Rosenblum, L. D. (1993). Visual influences on auditory pluck and bow judgments. *Perception & Psychophysics, 54*(3), 406–416.

Sams, M., Manninen, P., Suraka, V., Helin, P., & Kaettoe, R. (1998). McGurk effect in finnish syllables, isolated words and words in sentences: Effect of word meaning and sentence content. *Speech Communication, 26*(1–2), 75–87.

Schultz, R. T., Gauthier, I., Klin, A., Fulbright, R. K., Anderson, A. W., Volkmar, F., et al. (2000). Abnormal ventral temporal cortical activity during face discrimination among individuals with autism and asperger syndrome. *Archives of General Psychiatry, 57*(4), 331–340.

Schultz, R. T., Hunyadi, E., Conners, C., & Pasley, B. (2005). *fMRI Study of Facial Expression Perception in Autism: The Amygdala, Fusiform Face Area and Their Functional Connectivity.* Paper presented at the Annual meeting of the Organization for Human Brain Mapping, Toronto, CA.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26*, 212–215.

Summerfield, Q. (1979). Use of visual information for phonetic perception. *Phonetica, 36*, 314–331.

Surprenant, A. M., & Watson, C. S. (2001). Individual differences in the processing of speech and nonspeech sounds by normal-hearing listeners. *Journal of the Acoustical Society of America, 110*, 2085–2095.

Tager-Flusberg, H. (1981). On the nature of linguistic functioning in early infantile autism. *Journal of Autism and Developmental Disorders, 11*, 45–56.

Tager-Flusberg, H. (1982). Pragmatic development and its implication for social interaction in autistic children. In D. Park (Ed.), *Proceedings of the 1981 international conference on autism* (pp. 103–107). Washington, DC: NSAC.

Volkmar, F. R., & Lord, C. (1998). Diagnosis and definition of autism and other pervasive developmental disorders. In F. R. Volkmar (Ed.), *Autism and pervasive developmental disorders*. Cambridge: Cambridge University Press.

Williams, J. H. G., Massaro, D. W., Peel, N. J., Bosseler, A., & Suddendorf, T. (2004). Visual-auditory integration during speech imitation in autism. *Research in Developmental Disabilities, 25*, 559–575.